Transfer effects on spoken sentence comprehension and functional communication after working memory training in stroke aphasia

Lilla Zakariás, Christos Salis, Isabell Wartenburger

Author Note

Lilla Zakariás, Department of Linguistics, University of Potsdam, Potsdam, Germany; Christos Salis, Speech and Language Sciences, Newcastle University, Newcastle upon Tyne, United Kingdom; Isabell Wartenburger, Department of Linguistics, University of Potsdam, Potsdam, Germany

Correspondence concerning this article should be addressed to Lilla Zakariás, Cognitive Science, Department of Linguistics, University of Potsdam, Potsdam, Germany

E-mail: zakarias@uni-potsdam.de

Abstract

Recent treatment protocols have been successful in improving working memory (WM) in individuals with aphasia. However, the evidence to date is small and the extent to which improvements in trained tasks of WM transfer to untrained memory tasks, spoken sentence comprehension, and functional communication is yet poorly understood. To address these issues, we conducted a multiple baseline study with three German-speaking individuals with chronic post-stroke aphasia. Participants practised two computerised WM tasks (*n*-back with pictures and *n*-back with spoken words) four times a week for a month, targeting two WM processes: updating WM representations and resolving interference. All participants showed improvement on at least one measure of spoken sentence comprehension and everyday memory activities. Two of them showed improvement also on measures of WM and functional communication. Our results suggest that WM can be improved through computerized training in chronic aphasia and this can transfer to spoken sentence comprehension and functional communication in some individuals.

Keywords: aphasia, working memory, *n*-back training, transfer, sentence comprehension, verbal communicative abilities

1. Introduction

Individuals with aphasia (IWA) may present with concomitant cognitive deficits including deficits of short-term memory, working memory (WM)[1] (e.g., Friedmann & Gvion, 2003; Mayer, Mitchinson, & Murray, 2016; Nickels, Howard, & Best, 1997; Sung et al., 2009) and executive functions (e.g., Helm-Estabrooks & Albert, 1991; Nicholas, Hunsaker, & Guarino, 2017; Purdy, 2002; Zakariás, Keresztes, Demeter, & Lukács, 2013). WM is a complex cognitive construct referring to processes that support the temporary maintenance *and* manipulation of information (Baddeley, 2012; Engle, 2002; Martin, Kohen, Kalinyak-Fliszar, Soveri, & Laine, 2012). Manipulation in WM involves various processes, such as shifting attentional control between tasks or mental sets, updating and monitoring WM representations, inhibiting prepotent responses, and resolving different types of interference (Friedman & Miyake, 2004; Miyake et al., 2000). Such processes have been considered under the umbrella term executive functions (e.g., Miyake et al., 2000).

There is strong evidence suggesting that WM impairments can negatively influence various language processes in aphasia, such as lexical-semantic processing (Martin et al., 2012; Novick, Kan, Trueswell, & Thompson-Schill, 2009; Robinson, Blair, & Cipolotti, 1998), sentence comprehension (Novick et al., 2009; Sung et al., 2009; Wright, Downey, Gravier, Love, & Shapiro, 2007), spoken discourse and functional communication (Frankel, Penn, & Ormond- Brown, 2007; Fridriksson, Nettles, Davis, Morrow, & Montgomery, 2006; Keil & Kaszniak, 2002; Luna, 2011; Penn, Frankel, Watermeyer, & Russell, 2010; Ramsberger, 2005), and reading

[1] Short-term memory and WM are related constructs. It is generally acknowledged that short-term memory is responsible for the temporary maintenance and retrieval of information (Caplan & Waters, 2013), whereas WM is generally viewed as the combination of multiple components working together and actively manipulating information in short-term memory (Cowan, 2009). There is a multitude of theoretical accounts describing the relationship between short-term memory and WM. In the present paper we adopt the view that short-term memory is a component of WM (Baddeley, 2012; Cowan, 2009).

(Caspari, Parkinson, LaPointe, & Katz, 1998). Spontaneous recovery (Sharp, Turkheimer, Bose, Scott, & Wise, 2010) and responsiveness to language treatment have also been shown to relate to WM skills in aphasia (Brownsett, Warren, Geranmayeh, Woodhead, Leech, & Wise, 2013; Lambon Ralph, Snell, Fillingham, Conroy, & Sage, 2010).

With such strong links between WM and aphasia, researchers began to devise experimental treatments that heavily rely on WM, hypothesizing transfer of treatment effects to language functions. In these studies, treatments of WM included one or more WM tasks practised intensively, and treatment effects were measured on components of WM (i.e., near transfer) and language (i.e., far transfer), including spoken sentence comprehension (Eom & Sung, 2016; Francis, Clark, & Humphreys, 2003; Harris, Olson, & Humphreys, 2014; Salis, 2012; Salis et al., 2017; Zakariás, Keresztes, Marton, & Wartenburger, 2016), reading comprehension (Coelho, 2005; Mayer & Murray, 2002; Sinotte & Coelho, 2007), and spoken discourse (Paek & Murray, 2015; Peach, Nathan, & Beck, 2017). In the next section we discuss in detail treatment studies of WM and spoken sentence comprehension in people with non-progressive aphasia, which is the focus of the present paper.

1.1. Working memory treatments and sentence comprehension

Recent WM treatment studies that aimed to improve spoken sentence comprehension in aphasia reveal mixed findings, possibly due to substantial variations in participant characteristics, treatment tasks, intensity and duration of treatment, as well as variations in the domains and patterns of transfer detected. For example, Paek and Murray (2015) described a patient with mild anomic aphasia and semantic short-term memory deficit. The treatment included various tasks aiming to improve components of WM (i.e., updating, phonological loop) as well as semantic processing (see Table 1). The intervention was delivered remotely (teletherapy) consisting of 20 hourly sessions distributed over four weeks. Although the authors reported improvements in all training tasks, they observed near transfer effects only in one measure of short-term memory (identity span). With respect to far transfer, no substantial change was observed in spoken sentence comprehension. However, greater improvements were found in spoken discourse as measured by story-telling tasks. Additionally, improvements in short-term memory and spoken discourse were maintained at 6-week follow up.

Eom and Sung (2016) conducted a group study with six participants presenting with different types and severity of aphasia (see Table 1). They used a repetition-based treatment, incorporating sentences with varying length and syntactic complexity. The treatment combined repetition of sentences after auditory presentation, reconstruction of sentences by using word cards, and reading sentences aloud. Trained structures included active sentences with two- and three-argument verbs, passive sentences, conjoined sentences, and centre-embedded sentences with a subject-relative clause. Twelve sessions were administered over a month (three hourly sessions a week). With respect to the outcome, participants improved in the repetition of treated and untreated sentences, as well as in WM measured by digit and word span tasks. More importantly, they improved in the comprehension of treated syntactic structures (see Table 1).

Zakariás et al. (2016) used a computerised adaptive training approach (e.g., Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Novick, Hussey, Teubner-Rhodes, Harbison, & Bunting, 2014) utilising an *n*-back task with letters. *N*-back targets components of WM, such as updating (Cohen, Perlstein, Braver, Nystrom, Noll, Jonides, & Smith, 1997) and interference control (Kane, Conway, Miura, & Colflesh, 2007; Novick et al., 2014). The adaptive training task involved adjusting the difficulty level according to the participants' performance, ensuring that they always practised at an optimal level of difficulty. Training was delivered three to four times a week for a month (a total of 13 20-min sessions) to three Hungarian-speaking IWA (see Table 1). The authors detected a mixed pattern of training and transfer effects. One participant improved in the training task as well as untrained WM tasks and spoken sentence comprehension. Another participant improved in the training task and spoken sentence comprehension but did not show improvement in other measures of WM. The third participant did not show improvement in the training task but did show increases in performance, both in sentence comprehension and untrained WM. Zakariás and colleagues argued that individual differences in motivation as well as in cognitive abilities, such as interference control at the beginning of training could have influenced treatment outcome and transfer effects.

To replicate previous positive findings based on one IWA (Salis, 2012), Salis and colleagues (2017) delivered a training involving a recognition memory task (matching listening span) in five participants (for more information, see Table 1). The authors hypothesised far transfer to spoken sentence comprehension and improvements on psychosocial functioning as well as other communication skills after training. Participants received 27-30 treatment sessions. With respect to short-term memory (near transfer), changes were found only in one outcome measure (i.e., digit matching listening span). None of the changes observed in spoken sentence comprehension was statistically significant (see Table 1). As for the psychological measures of communication, a statistically significant increase was observed only in case of one participant.

In summary, although previous results suggest that components of WM indeed can be flexibly improved with training, the extent of transfer to untrained abilities and its boundary conditions are not well understood. There have been variations in the domains (i.e., WM and/or language abilities) and patterns of transfer detected after training: some researchers reported substantial effects on WM (e.g., Eom & Sung, 2016, but for null effects, see Salis et al., 2017), spoken sentence comprehension (e.g., Eom & Sung, 2016; Salis, 2012; Zakariás et al., 2016, for null effects, see Paek & Murray, 2015), and spoken discourse (Paek & Murray, 2015), whereas others did not find any effects on untreated processes after training (Salis et al., 2017). Although the role of WM in syntactic comprehension has drawn much attention in the last decades (e.g., Caplan, Michaud, & Hufford, 2013; Caplan & Waters, 2013 for review; Fedorenko, 2014; Haarmann, Just, & Carpenter, 1997), only Eom and Sung (2016) has investigated the specificity of transfer effects on syntactic comprehension. The inconsistent pattern of transfer can be observed also across participants: for example, in Zakariás et al. (2016), some participants showed near but not far transfer effects, while others showed the opposite pattern. In addition, there is limited knowledge as to which participants – with respect to type and severity of aphasia or degree of impairment in certain linguistic and WM processes – can benefit from training. Although some researchers suggested that training WM might bear a higher potential for IWA with moderate or severe sentence comprehension deficits (e.g., Salis, 2012; Zakariás et al., 2016), Eom and Sung (2016) concluded that WM treatments might be more beneficial for people with relatively preserved comprehension abilities. Based on observations that IWA with WM spans of zero performed at chance on the sentence comprehension tasks, whereas participants with WM spans of 1 or 2 showed normal performance on the tasks, Caplan et al. (2013) suggested that there is a minimal WM capacity (span above 1) that is needed to perform normally in sentence comprehension. These findings also suggest that WM treatments might bear a higher potential for IWA demonstrating with severe WM impairments.

In summary, potential training and transfer effects following WM training in aphasia warrant further systematic study to refine our understanding of the nature and the underlying mechanisms of transfer of WM training to different levels of linguistic processing.

Insert Table 1 here

1.2. Extending the ecological validity of WM trainings in aphasia: motivation, functional communication, and everyday memory

Besides resolving the issues discussed above, the present study aimed to extend the investigation to motivation and two relevant domains of target in aphasia. Research from other literature domains, beyond aphasia, suggests that motivation plays a substantial role in the effectiveness of WM training (Jaeggi, Buschkuehl, Jonides, & Shah, 2011; Jaeggi, Buschkuehl, Shah, & Jonides, 2014; Katz, Jaeggi, Buschkuehl, Stegman, & Shah, 2014). Studies using *n*-back tasks for training in healthy children (Jaeggi et al., 2011) and healthy young adults (Jaeggi et al., 2014) suggest that motivational factors, such as interest in or engagement with the training activity mediates improvement in the training task, and, in turn, transfer to other untrained abilities (Lindeløv et al., 2016). Yet, motivation is an overlooked aspect of training, and to our knowledge no study has yet incorporated measures of motivation in WM treatment studies in aphasia.

For most IWA, the important goal of linguistic rehabilitation is improvement in functional communication, that is, the individual's ability to understand and convey information in everyday life situations (Blomert, Kean, Koster, & Schokker, 1994; Lind, Kristoffersen, Moen, & Simonsen, 2009). Therefore, such improvements are seen as the gold standard for demonstrating the effectiveness of any intervention. Despite its importance in aphasia rehabilitation and the suggested link between WM and functional communication (Frankel et al., 2007; Fridriksson et al., 2006; Keil & Kaszniak, 2002; Luna, 2011; Penn et al., 2010; Ramsberger, 2005), only very few studies have investigated transfer effects after WM training on functional communication (Murray, Keeton, & Karcher, 2006; Salis et al., 2017).

Although aspects of memory functioning in everyday life activities, such as difficulty in remembering appointments or recognizing familiar faces have been observed after stroke (e.g., Stewart, Sunderland, & Sluman, 1996; Wilson, Cockburn, Baddeley, & Hiorns, 1989), studies have provided limited or incomplete information about participants' aphasia. For instance, the presence and the number of IWA in some stroke studies are unclear (e.g., Barker-Collo, Feigin, Parag, Lawes, & Senior, 2010), or the diagnostic method to identify aphasia is not described (e.g., Duffin, Collins, Coughlan, O'Neill, Roche, & Commins, 2012). Thus, knowledge about the extent of everyday memory problems, recovery of everyday memory, and its improvement in response to treatment in participants presenting with aphasia is limited (for the only study see Vallat-Azouvi, Pradat-Diehl, & Azouvi, 2014).

1.3. The present study

In the present study, we used the *n*-back task for the training. *N*-back is a complex WM task involving multiple processes, such as encoding incoming stimuli, monitoring, maintaining, and updating WM representations, establishing and maintaining bindings between memory contents and their temporal context, as well as resolving interference between WM representations (Kane et al., 2007). In a typical *n*-back task, participants are presented with a continuous stream of items and are instructed to judge whether an item matches a previous one that was presented *n* items (e.g., $n = 1$, $n = 2$) before. Although the task commonly used to investigate WM in language-impaired populations, results regarding its reliability in aphasia are mixed with some studies showing excellent test-retest reliability (Mayer and Murray, 2012), whereas others showing only acceptable test-retest reliability (Zakariás et al., 2016). Varying test-retest reliabilities are likely due to differences in task stimulus materials, procedures, and the measures used to describe performance, as well as participants' cognitive and linguistic profile (cf., DeDe, Ricca, Knilans, & Trubl, 2014). Despite such challenges, certain properties of the task enhance research validity and treatment fidelity (i.e., the reliability of the administration of an intervention) in studies using *n*-back as a training task in aphasia. These properties include, among others, that the task does not require speech response, or that the task structure is easy to convey and the administration is simple and in most cases automatized.

The present study was motivated by the need to strengthen and extend the evidence base of WM treatments in aphasia and also improve our knowledge as to why inconsistent patterns of transfer were reported in previous studies. Our main objective was to systematically investigate patterns and potential domains of transfer after WM training. To this end, we chose a set of outcome measures that allowed for a systematic investigation of potential transfer effects, ranging from the training task (*n*-back) to very far transfer (functional communication). To assess the specificity of transfer effects and to better understand the underlying mechanisms of transfer on sentence comprehension, our outcome measures included specific syntactic structures that have been proposed to involve WM processes (e.g., non-canonical structures with varying complexity; Caplan et al., 2013; Haarmann, Just, & Carpenter, 1997). In addition, we aimed to extend earlier reports of WM training related transfer effects in aphasia by extending the ecological validity of our findings. Therefore, we included a set of far transfer tasks that covered a broad range of WM-relevant language and everyday functions, such as spoken sentence comprehension, functional communication, and everyday memory. To capture the effects of motivational factors on training performance across time, we monitored participants' motivation on a daily basis. In summary, the research questions in this study are:

(1) Does WM training transfer to cognitive domains targeted by the training but measured by untrained tasks (i.e., near transfer) in IWA?

(2) Does WM training transfer to spoken sentence comprehension, functional communication, and everyday memory (i.e., far transfer) in IWA?

(3) Are training and transfer effects maintained over time (i.e., at 4-6 weeks follow up)?

(4) Do motivational factors play a role in IWAs' WM training performance?

Our design followed an earlier report by Zakariás and colleagues (2016), that suggested that intensive *n*-back training can lead to improvements on untrained WM domains and spoken sentence comprehension (i.e., near and far transfer, respectively) in aphasia. We expected that IWA improving on the training tasks will improve on all outcome measures, but not on the non-targeted control measure (oral word reading). In addition, we hypothesized that stable and generally high interest levels (i.e., a factor of motivation) would be associated with greater improvement in the training task.

2. Methods

2.1. Participants

Three IWA participated in the study. Participants were included based on the following criteria: (1) aphasia as a result of left hemisphere stroke, (2) at least eight months post-onset, (3) German as the native language, (4) self-reported pre-stroke right-handedness, (4) moderate to severe impairment in sentence comprehension together with good single word comprehension (based on the Aachen Aphasia Test, AAT, Huber, 1983), (5) a score of three items or below in a verbal WM task (i.e., listening span, developed based on Tompkins, Bloise, Timko, & Baumgaertner, 1994[2]) and a score of five items or below in a computerised visuo-spatial WM task (Corsi block tapping). Exclusion criteria were: (1) bilateral lesions, (2) additional neurological or psychiatric disorder, and (3) participation in speech and language therapy during the time of study. Participants were recruited through the aphasia database of the University of Potsdam.

Participant 1 (E.Q.) was a 39-year-old female six years post-onset. She worked as a beautician at the time of her stroke. Prior to the study, she had received individual speech and language therapy, which was suspended during the present study (altogether for four months). Participant 2 (I.B.) was a 77-year-old female 25 years post-onset. She had studied German literature and history, then had worked as a teacher, and later as a television editor. At the time of the study she was retired, was living with her husband and was not participating in any therapy. Participant 3 (M.N.) was a 51-year-old female 15 years post-onset. Her right hand and arm were still non-functional at the time of the

[2] The procedure of the listening task followed that of Tompkins et al., 1994. Stimuli were modified to make the task suitable for use with participants with aphasia. For stimuli and procedure of the task see supplementary material – Table S3.

study. Before the stroke, she had worked as a trained nurse. She was not participating in speech and language therapy but received physiotherapy once a week during the present study. The study was approved by the local research ethics committee of the University of Potsdam. The participants provided informed voluntary consent during the initial meetings. There was no dropout. Participants' biographical information and initial scores on various assessments are shown in Table 2.

Insert Table 2 here

2.2. General design and procedures

For each participant, a multiple-baseline (with control) experimental design was adopted (see Figure 1 for an overview). Each participant received two blocks of WM training: (A) a visual *n*-back task with pictures; and (B) an auditory *n*-back task with spoken words. Following random assignment of participants to block order, E.Q. and M.N. started with the visual WM training, followed by the auditory WM training. I.B. received the training in the reverse order (auditory WM training, followed by visual WM training).

Participants were assessed before the first training block (i.e., pretest) and after the second training block (i.e., posttest) on several memory and language tasks. Assessments were distributed over six sessions in both test phases. The experimental tasks were administered twice in both test phases. In addition, four to six weeks after completion of the posttest, we conducted one follow-up test session –including a subset of the tasks administered at pre- and posttest – to tap into the time-course of training induced changes and maintenance of potential transfer effects. Experimental tasks were administered once at follow-up. The training blocks consisted of eight sessions each (approximately 25-35 minutes/session, three-four sessions/week), resulting in a four to five-week total training period. After each training session, participants completed a motivation questionnaire assessing their subjective experience related to the training. Altogether, the study consisted of 30 sessions for each participant, lasting approximately 10 weeks (see Figure 1). All computerised tasks were delivered by Presentation® software (Version 18.3) on a Lenovo X201 ThinkPad® (E.Q.) or a Lenovo IdeaPad U310 (I.B. and M.N.).

Insert Figure 1 here

2.2.1. Training tasks

Based on Zakariás et al. (2016), we created two *n*-back tasks with identical design and procedure (one with pictures, one with spoken words). The two *n*-back tasks were chosen to tax verbal short-term memory as well as domain general executive functions (e.g., interference control) (Redick & Lindsey, 2013; Kane et al., 2007). Since the participants' word comprehension abilities were relatively good at the beginning of the training, we supposed that both semantic and phonological short-term memory would be activated, at least to some extent, in both tasks.

Stimuli. Eight stimuli sets, each including eight stimuli belonging to different semantic categories (64 stimuli altogether), were created for the eight blocks in both training tasks (pictures, words). This allowed us to present eight stimuli belonging to different semantic categories in each block. For the *n*-back with pictures, eight pictures from eight semantic categories (animals, furniture, clothes, body parts/tools, vehicles/musical instruments, food, toys, home) were taken from the coloured version

(Rossion & Pourtois, 2004) of the Snodgrass and Vanderwart (1980) set. When there were no eight items belonging to the same category available, we chose the remaining items from another category (e.g., vehicles and musical instruments, respectively). For the *n*-back with spoken words, eight words from the eight semantic categories (animals, vegetables/drinks, vehicles, furniture, musical instruments/toys, tools, clothes, professions) were recorded by a female native German speaker in an acoustically shielded recording studio, at a sampling rate of 44.1 kHz (16 bit, mono). The speaker was instructed to produce the words naturally with normal intonation and speech rate. Auditory recordings were created, edited, denoised, and segmented into single word sound files using Audacity®2.1.2. Words across the blocks were balanced for length (i.e., each block included three 1-syllable and five 2-syllable words) as well as for lexical frequency (i.e., no significant difference between the blocks). Frequency values were obtained from the CLEARPOND database (Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities, Marian, Bartolotti, Chabal, & Shook, 2012). Any two words in a block were checked not to result in a meaningful compound word if presented one after the other by a native German speaker.

Note that the limited number of pictures available in the Snodgrass and Vanderwart (1980) database as well as the limited number of words meeting the criteria in our auditory *n*-back did not allow us to choose items belonging to the same eight categories in both tasks. Due to the category, frequency, and length constraints, 22% of the stimuli overlapped between the two training tasks.

Procedure. Participants were exposed to a continuous stream of stimuli (i.e., either pictures or spoken words) and were asked to press a button on the keyboard when the stimulus presented was the same as the one that had been presented in *n* preceding trials (see Figure 2). In addition, "lures" were incorporated into the task; stimuli that were the same as the one presented *n-1* or *n+1* (but not *n*) trials before, requiring participants to resolve the conflict between the representation of the target and that of a highly familiar lure. The tasks were adaptive, that is, the task difficulty was always continuously adapted according to participants' performance by means of automatic computer algorithms. If a given threshold (described below) was reached at the end of a block, then difficulty level for the upcoming block automatically increased by one, if the threshold was not reached for four consecutive blocks, the difficulty level decreased by one. Increase in difficulty level meant advancing through three lure levels at each value of *n* (i.e., no lures, *n+1* lures only, and both *n+1* and *n-1* lures), then advancing through to the next *n*.

The required threshold was defined based on three measures: (1) hit rates (proportion of responses to targets), (2) false alarm rates for non-targets (proportion of responses to non- targets), and (3) false alarm rates for lures (proportion of responses to lures), when lures were present in the block. The threshold was defined as having a hit rate above or equal to 80%, a false alarm rate for non-targets below 30% (E.Q. and I.B.) or 10% (M.N.)[3], and a false alarm rate for lures (when lures were present in the block) below 10%. In the *n*-back with pictures, stimuli were presented sequentially on a computer screen at a rate of 3 seconds (stimulus length: 1500 ms; interstimulus interval: 1500 ms) per trial. In the *n*-back with spoken words, stimuli were presented at the same rate (mean stimulus length: 785 ms, range: 445-1180 ms) via a loudspeaker (Speedlink Ellipz Stereo Speakers). Volume was adjusted to each participant's comfort

[3] Because M.N.'s false alarm rate for non-targets was very high (above 20%) in blocks of the first training session (Training A), keeping the threshold for false alarms at 30% for the whole time of training would have let her advance to the next levels without actually mastering the task (based on trial-by-trial strategy). Therefore, after the first session we changed it from 30% to 10% for her.

with the volume control on the loudspeaker. Participants responded manually by pressing the SPACE bar on the computer keyboard. No responses were required for non-target items. One training session comprised six to eight blocks consisting of $18 + 5*(n - 1)$ trials including 5 targets, resulting in a daily training time of 25-35 minutes. The number of lures in blocks including lures was always five. The sequence of the stimuli in each block was randomized in both tasks.

Feedback. Recent studies have shown that feedback can impact participants' performance during training as well as keep them engaged with the training regimen (Jaeggi et al., 2011; Katz et al., 2014). To maximise motivation and compliance with the training, participants received three types of feedback during training. The first type of feedback was provided after each block. It was based on participants' hit rate, false alarm rate for non-targets, and false alarm rate for lures, by displaying their average performance in percentage on the screen. The second type of feedback was displayed based on the pattern of participants' errors. When the false alarm rate for non-targets was higher than 50%, they were given the feedback, "Caution: you might be pressing the button too often." When the false alarm rate for non-targets was below 50%, but false alarm rate for lures was above 60%, the feedback was, "Caution: there are some tricky trials that might lure you into pressing the button." If hit rates were below 40%, the feedback was, "Caution: you're pressing the button quite rarely." The third type of feedback was provided after certain trials: after each hit and at 80% of the misses, a message was displayed on the screen ("Good!" and "Missed out!", respectively). The first and second types of feedback were always also read aloud to the participants by the trainer.

Delivery. The training was delivered in the participants' home in a quiet room. The training to E.Q. was delivered by a trained speech-language pathologist (SLP) and a SLP student (Student 1). The training to I.B. and M.N. was delivered by two SLP students (Student 2 and 3, respectively). All trainers had completed the same 3-hour training session regarding conducting and administering the training tasks (i.e., setting up the computer and the tasks, providing the computer-generated instructions and feedback to the participants, saving data on computer). The trainer was present at all training sessions.

Insert Figure 2 here

2.2.2. Outcome measures

2.2.2.1. WM 1: N-back with letters (near transfer)

This experimental task was used to assess near transfer effects across stimuli. Because the structure of this task was the same as those of the training tasks but the stimuli were different, it allowed us to assess task-specific effects. Participants were exposed to a stream of letters. One letter appeared on each trial and participants had to respond by pressing the SPACE bar on the keyboard when the stimulus presented was the same as the one presented n trials before. We varied n within subjects, and all participants performed the n-back task first with $n = 1$, then with $n = 2$, and finally with $n = 3$. In all conditions, the task consisted of three blocks, with 90 trials (including 15 targets) in total. In addition, a practice block with 15 trials (including 3 targets) was also included with $n = 1$. Data of the practice block was not included in the analysis. Blocks were separated by self-paced resting periods. For each trial, a letter was sampled from the same pool of letters (i.e., B, F, K, H, L, S, C, and N). Sampling was pseudorandomized to always provide exactly five targets in all blocks. In each trial, the letter was presented in the middle of the screen for 1500 ms, and trials were separated by a 1500 ms interstimulus interval (temporal parameters in the task were the same as in the training tasks). No feedback was provided to the participants.

2.2.2.2. WM 2: Running span (near transfer)

This experimental task was used to assess transfer effects on updating (Pollack, Johnson, & Knaff, 1959; Collette et al., 2007). Running span involves similar processes as the *n*-back task, but has a different structure (Collette et al., 2007). Because it was not practised during the training, it also allowed us to separate task-specific from process-specific effects. The running span task was adapted to aphasia as follows: in each trial, participants were presented with a list of digits (one digit at a time), and were asked to respond by pointing the *n* last number of digits (*n*-span) when the list presentation ended. Importantly, participants were informed about *n* (i.e., how many digits they would need to report), but not the list length (i.e., they did not know when the list would end). Lists included two-six random digits (from the set 1-9) presented via computer. Digits appeared in the middle of the screen, one at a time, for 1500 ms. Immediately after each list, digits (separated by commas) together with one blank per to-be-recalled item appeared on the screen. For example, given the list, “6, 2, 4, 7, 5” in a 2-span condition, participants saw “6, 2, 4, _, _”. Participants had to report digits by pointing to the corresponding digits on a sheet of paper, which was positioned in front of them on a desk (i.e., no oral response was required). The experimenter noted down the answers on a scoring sheet and pressed ENTER to advance to the next trial. The task comprised three blocks of five trials (altogether 15 experimental trials), with span increased from 1 to 3 across blocks. Experimental trials were preceded with two probe trials with 1-span length. Probe trials were not included in the analysis. The number of correct trials was calculated in the task (max. 15).

2.2.2.3. Spoken sentence comprehension 1: TROG-D (far transfer)

This standardized test measures the spoken comprehension of grammatical structures (Fox-Boyer, 2013). It comprises 84 multiple-choice items, organized into 21 blocks, each testing a different grammatical structure. The grammatical complexity and hence difficulty of the linguistic material increases with each block. For each item, an array of four coloured pictures is presented and the task is to select the picture matching the word, phrase or sentence read aloud by the experimenter. For each item, there are three – either lexical or grammatical – distractor pictures and one target picture. Each participant completed the entire test in approximately 30 minutes. We calculated and analysed the number of correct responses in the task.

2.2.2.4. Spoken sentence comprehension 2: Token test (far transfer)

This standardized test measures comprehension of spoken commands of increasing length and, in the last subtest, of increasing grammatical complexity (Huber, 1983). Understanding of commands requires pointing to or manipulating with plastic tokens with different sizes, shapes, and colours. This version taken from the AAT (Huber, 1983) consists of five subtests, including 10 sentences in each. The number of correct responses was the dependent variable (max. 50).

2.2.2.5. Spoken sentence comprehension 3: Sätze verstehen (far transfer)

This standardized test measures the comprehension of spoken sentences varying in syntactic complexity, semantic reversibility, and length (Burchert, Lorenz, Schröder, De Bleser, & Stadie, 2011). It consists of 204 sentences and uses a sentence-picture matching paradigm with two or four pictures (irreversible and reversible sentences with two-argument verbs, and relative clauses, respectively). It includes *short* and *long irreversible sentences* (22 sentences each), *case-marked canonical* (SVO) and *non-canonical* (OVS) *reversible sentences* (20 sentences each), *number-marked canonical* (SVO) and *non-canonical* (OVS) *reversible sentences* (20 sentences each), and *right-branching* and *centre-embedded subject* and *object relative clauses* (20 sentences each). Each participant completed the entire test over three sessions. With the inclusion of specific syntactic structures, the test assesses the specificity of transfer effects in terms of underlying mechanisms of transfer on sentence comprehension. The number of correct responses was calculated for each syntactic structure. In addition, aggregate scores in the canonical (i.e., SVOs plus SRCs) and the non-canonical (i.e., OVSs plus ORCs) conditions, as well as a total score (i.e., the number of all correct responses in the task) were calculated.

2.2.2.6. Functional communication: Amsterdam-Nijmegen Everyday Language Test, ANELT (far transfer)

This test measures spoken communicative skills (Brunner & Steiner, 1994). There are two parallel versions (ANELT 1 and 2), each consisting of 10 items involving familiar everyday life situations (e.g., calling a doctor, talking to a cashier). Items are presented verbally to the participant. The experimenter is instructed to avoid conversing with the participant but to act as an interested listener, while the participant answers the items as a brief monologue. The administration of the ANELT is recorded on audiotape for later scoring and it takes 15-20 minutes to administer. The response of the participant for each item is rated on two 5-point scales (0-4), evaluating the *understandability of the message* and the *intelligibility of the utterance (sic)* (scale A and B, respectively). The maximum score for both understandability and intelligibility is 40.

Finally, we performed a quantitative analysis of the data (Nicholas & Brookshire, 1993). Language samples were transcribed and analysed for number of words, number of correct information units (CIUs), the percentage of correct information units (%CIUs), and efficiency (e.g., CIUs/minute, words/minute). A speech and language pathology student previously trained in clinical and experimental linguistics completed the transcription of the speech samples. For information on scoring the scales and analysing word and CIU measures, see the Data analysis section.

2.2.2.7. Everyday memory questionnaire (EMQ, far transfer)

We adapted the everyday memory questionnaire developed by Sunderland, Harris, and Baddeley (1983) to aphasia. Thirty-one examples of memory difficulties were included in the present questionnaire (see supplementary material – Table S1). A close relative or partner of the participants was asked to judge how often a difficulty occurs in the participant's activities of daily living, using a 5-point rating scale (where 0 indicates *never* and 4 indicates *quite often*). Difficulties were grouped under the headings "Speech" (e.g., "Confusing the names of common things or using the wrong names"), "Faces and places" (e.g., "Forgetting where s/he has put something, losing things around the house"), "Actions" (e.g., "Forgetting to do some routine thing that s/he would normally do once or twice in a day"), and "Learning new things" (e.g., "Unable to pick up a new skill such as a game or working some new gadget after s/he has practiced once or twice”). Items followed each other in a fixed order. Ratings for each type of memory difficulty were summed and analysed.

2.2.2.8. Control task: Oral word reading

As oral word reading potentially does not tax WM majorly, we used it as a control task to test that possible improvements on the outcome measures were specifically related to the WM training. The task consisted of simple words (25 items) and compound words (20 items) with varying length (1-4 syllables) and frequency (low frequent vs. high frequent), as well as one-syllable pseudo-words (15 items). We selected words from Lorenz, Heide, and Burchert (2014) and pseudowords from the subtest of LeMo 2.0 (Stadie, Cholewa, & De Bleser, 2013). Items were printed separately on A4 format paper sheets (font size 44) and presented one at a time in a fixed order. Participants were instructed to read aloud the words, each within a 10 seconds limit. If there was no response within this time limit, the examiner proceeded to the next item. The task took approximately 10 minutes. The participants’ responses were audio recorded and later transcribed and scored by two SLP students (one of them previously mastered in clinical and experimental linguistics). The total number of correctly read items was calculated.

2.2.3. Motivation questionnaire (MQ)

We developed a self-report motivation questionnaire based on Jaeggi et al. (2011) and McAuley, Duncan, and Tammen (1989). The questionnaire consisted of 10 questions assessing the participants’ *interest/enjoyment*, *perceived competence*, and *effort/importance* while performing the training. Participants responded on a 7-point Likert scale from 1 (low degree of approval) to 7 (high degree of approval). Four questions focused on interest/enjoyment (e.g., "How much did you enjoy the activity today?" – 1: *not at all*, 7: *a lot*), three questions on perceived competence (e.g., "How satisfied are you with your performance today?" 1: *not satisfied at all*, 7: *very satisfied*) and three on effort/importance (e.g., "How much effort did you put into this today?" – 1: *nothing at all*, 7: *a lot*, see supplementary material – Table S2). Participants completed this questionnaire after each session. Experimenters were instructed to read aloud the questions and note the response of the participant. They were also instructed to explain questions if needed but to avoid influencing the participants' response selection in any way. We calculated the mean score for each factor for each session to capture the changes in motivation over time and possibly relate them to the performance pattern in the training tasks.

Similar to the training sessions, test sessions were conducted by an SLP and by SLP students. The same person(s) for each participant conducted test and training sessions. Importantly, for outcome measures that were obtained by scoring/rating the participant's responses by the experimenter (i.e., that were not computer generated) the responses were also scored by an independent experimenter and tested for inter-rater reliability (for details of this step, see the Data analysis and Results section). All experimenters participated in a 5×2 hour training provided by the first author of the paper regarding the conduction, administration, and scoring of the tasks.

2.3. Data analyses

2.3.1. Performance change in the training tasks and outcome measures

Individual performances on the training tasks were tested using non-parametric Spearman correlations. Based on Vallat, Azouvi, Hardisson, Meffert, Tessier, and Pradat-Diehl (2005), Fisher's exact test was used to compare performance in the two baselines (to demonstrate stability, p should > .1, two-tailed).

We used Fisher's exact and McNemar's test (p < .05, one-tailed) to compare performance on pretest and posttest, by taking the aggregate data obtained on two occasions for both pretest and posttest (note that data was obtained on two occasions only in the experimental tasks). To investigate long-term maintenance of potential effects (i.e., compare performance between posttest and follow up), we used Fisher's exact and McNemar's chi square test. Group level performance on the outcome measures was analysed with Wilcoxon signed rank test (p < 05, one-tailed). The relationship between the improvement in the training task and changes in motivation factors was tested with Spearman correlation (*rho*).

2.3.2. Inter-rater reliability

Inter-rater reliability represents the correspondence between raters' scores, thus indicates a measure of reliability for the collected data (Morgan & Morgan, 2008). Inter-rater reliability of the sentence comprehension tests and the running span (i.e., in case of dichotomous data) was examined using proportion scoring agreement, by dividing the number of agreements by the number of agreements plus disagreements (Franklin, Allison, & Gorman, 2014; Morgan & Morgan, 2008). The running span and the sentence comprehension tests were scored on 55% of the samples (range 33-75%) by two experimenters who were both present during the assessment (i.e., the trainer and an independent but not blind assessor). Inter-rater reliability of the oral tasks' measures was determined using an ICC two-way random effects model (ICC(2,k)) (Franklin et al., 2014). The oral tasks (i.e., ANELT and word reading) were audiotaped and transcribed; 100% of the oral word reading, 33% and 17% of the ANELT speech samples (qualitative and the quantitative analysis, respectively) were analysed by two independent experimenters who were also blind to the study phase.

3. Results

Inter-rater reliability was excellent for all tasks: mean proportion scoring agreement was 1 for the running span, .98 for the Token, .98 for the Sätze verstehen, and 1 for the TROG-D. ICC(2,k) was .96 for the oral word reading, .74 for the ANELT (Scale A), and ranged between .85-.99 for quantitative measures of the ANELT. All discrepancies were resolved by consensus prior to analysis.

Participants demonstrated unstable baseline in some conditions: E.Q. and M.N. were not stable in the 3-back condition of the *n*-back with letters (Fisher's exact test, $p = .042$ and $p = .035$, respectively). I.B. was not stable in the running span (Fisher's exact test, $p = .042$). To get a more accurate picture of the participants' performance, we took the aggregate data obtained on two occasions for both pretest and posttest in the tasks.

3.1. Training tasks

To analyse performance at the individual level (Figure 3), correlations between number of training sessions and mean difficulty level at a session were calculated using Spearman correlation coefficient. I.B. showed a significant increase in performance in both the auditory and the visual training ($rho = 1$, $p < .01$ and $rho = .786$, $p < .05$, respectively), whereas E.Q. and M.N. only improved in the first training comprising the visual *n*-back task ($rho = .905$, $p < .01$ and $rho = 1$, $p < .01$, respectively).

Comparisons between posttest and follow up revealed changes in the participants' performance over time. With respect to the *n*-back with pictures, E.Q. showed a significant increase in performance in 2-back (Fisher's exact test, $p = .045$) and I.B. showed a significant decrease in performance in 3-back (Fisher's exact test, $p = .001$). With respect to the *n*-back with spoken words, E.Q. showed a tendency for a decrease in performance in the 2-back condition (Fisher's exact test, $p = .085$). In summary, participants consistently showed performance increases during training. However, improvement was not consistently maintained until 6-weeks after posttest.

Insert Figure 3 here

3.2. Outcome measures

Overview of the results of the outcome measures is in Table 3.

3.2.1. WM 1: N-back with letters

Aggregated scores showed that E.Q. improved significantly in 2-back and 3-back (Fisher's exact test, $p = 03.$ and $p < .001$, respectively), I.B. improved significantly in 2-back and 3-back (Fisher's exact test, $p = .024$ and $p = .034$, respectively), whereas M.N. did not improve in any of the conditions. Group level analysis showed no significant improvement in any of the conditions ($p > .05$ for all conditions). Note that in the 1-back condition E.Q. and I.B. were close to ceiling already at the beginning of the training.

3.2.2. WM 2: Running span

Analysis of the number of correct trials showed that none of the participants improved in the running span task (Fisher's exact test, $p > .05$ for all participants). Group level analysis showed a tendency level improvement in the task ($Z = -1.60$, $p = .054$)

3.2.3. Spoken sentence comprehension 1: TROG-D

M.N. significantly improved between pretest and posttest (McNemar chi square = 5.281, $p = .011$) and the improvement was maintained also at follow up (comparing posttest and follow up: McNemar chi square, $p > .1$); I.B. showed a tendency level improvement (McNemar chi square = 3.6, $p = .054$) between pretest and follow-up; whereas E.Q. did not improve. Group level analysis on total scores showed a tendency level improvement between pretest and follow-up ($Z = -1.60$, $p = .054$) as well as posttest and follow-up ($Z = -1.34$, $p = .09$). Thus, we detected a tendency for improvement on the comprehension of grammatical structures coupled with heterogeneous individual performance patterns.

3.2.4. Spoken sentence comprehension 2: Token test

Comparing pretest and posttest performance, a tendency towards improvement was found for E.Q. and M.N. (McNemar chi square = 2.37, $p = .061$ and chi square = 2.207, $p = .068$, respectively), whereas no significant change in performance was found for I.B. Group level analysis showed a tendency level improvement in the task ($Z = -1.41$, $p = .07$).

3.2.5. Spoken sentence comprehension 3: Sätze verstehen

E.Q. significantly improved in the comprehension of number-marked OVS sentences (McNemar test chi square = 7.53, $p < .01$) and non-canonical structures (McNemar chi square = 6.618, $p < .01$); I.B. significantly improved in the comprehension of canonical structures (McNemar chi square = 8.33, $p < .01$) and showed a tendency for increase in the total score (McNemar chi square = 1.75, $p = .09$); whereas M.N. did not improve in any of the conditions. At group level they showed a tendency for increase in the comprehension of right-branching subject relative clauses (Z = -1.34, $p = .09$) and centre-embedded object relative clauses (Z = -1.60, $p = .054$), and in the total score ($Z = -1.60$, $p = .054$).

3.2.6. Functional communication: ANELT

Analysis of the understandability scores (scale A) showed a significant positive change in I.B.'s functional communication ($U = 16.5$, $p < .01$). E.Q. and M.N. also showed an increase in performance but these were not statistically significant. Group level analysis showed a tendency level improvement in the task ($Z = -1.60$, $p = .054$).

Analysis of quantitative measures complemented this picture: M.N. significantly improved in number of words ($U = 17$, $p < .01$) and CIUs ($U = 19$, $p < .05$), I.B. significantly improved in percentage of CIUs ($U = 19$, $p < .05$) and showed a statistical tendency for improvement in CIUs/min ($U = 25$, $p = .056$), whereas E.Q. did not show statistically significant improvement in the task. At group level they showed a tendency level increase in performance according to the CIUs, %CIUs, and CIUs/min ($Z = -1.60$, $p = .054$ for all three measures).

3.2.7. Everyday memory questionnaire

Ratings for each type of memory failure were summed. We only analysed the total score in the section 'Speech' for each participant and the total score in the section

‘Learning new things’ for M.N., because in the other sections there was virtually no error reported. Scores in ‘Speech’ showed a tendency level decrease in memory failures for E.Q. and I.B. ($Z = -1.53$, $p = .063$ and $Z = -1.41$, $p = .078$, respectively) but a significant increase in memory failures for M.N. ($Z = -1{,}90$, $p = .028$). Scores in ‘Learning new things’ showed a statistically significant decrease in memory failures for M.N. ($Z = -1.73$, $p = .041$).

Insert Table 3 about here

3.2.9. Control task: Oral word reading

Pre-post comparisons for oral word reading showed that the participants’ performance remained stable on the task (Fisher’s exact test, $p > .05$ for all participants).

3.3. Motivation questionnaire

Mean scores were calculated for each motivation factor (i.e., interest/enjoyment, perceived competence, and effort/importance), based on each participant’s ratings that were elicited in each session. Changes in the motivation scores were analysed on a descriptive basis as well as statistically compared to the changes in performance on the training tasks by means of Spearman rank correlation coefficient. Changes in each motivation factor can be seen in Figure 4 for each participant.

E.Q. and I.B. reported moderate to high interest in the training tasks; their interest levels remained stable throughout the training. Both participants put great effort into the training tasks over the whole training period. M.N, however, showed a considerable fluctuation in all motivation factors. She reported greatly varying interest levels, with a decreasing tendency in the second training block. In addition, she reported generally lower effort levels than the other two participants during the whole training period.

For M.N., changes in perceived competence were significantly associated with changes in performance in the second training block (*rho* = .89, *p* = .007), suggesting that she was able to evaluate her performance on the training task. For E.Q., changes in effort were significantly associated with changes in performance in the first training block (*rho* = -.817, *p* = .025), suggesting that the more effort she put into the training task, the more she improved. All the other comparisons between changes in motivation factors and in performance on the training tasks were not statistically significant.

Mean interest and perceived competence scores showed a positive correlation (at the level of tendency) both in the first and the second training block (*rho* = .67, *p* = .068 and *rho* = .67, *p* = .097, respectively) for I.B., a positive correlation at the level of tendency in the first training block (*rho* = .66, *p* = .078) and a significant positive correlation in the second training block (*rho* = .96, *p* < .001) for M.N., and a tendency level negative correlation in the second training block for E.Q. (*rho* = -.66, *p* = .073). In addition, mean effort scores showed a significant positive correlation with mean interest scores and a tendency for a positive correlation with mean perceived competence scores (*rho* = .852, *p* = .007 and *rho* = .66, *p* = .076, respectively) in the first training block for E.Q.

Insert Figure 4 here

4. Discussion

In this study, we investigated whether WM training effects transferred to unpractised WM and spoken sentence comprehension tasks, as well as to functional communication and everyday memory. The training targeted different components of WM, such as maintaining and updating WM representations and interference control. Consistent with previous results in related studies (e.g., Eom & Sung, 2016; Paek & Murray, 2015), participants showed improvements in the training tasks. However, the patterns of improvement were not consistent across the two training blocks (i.e., two participants improved only in the first block comprising the *n*-back with pictures). Performance patterns suggest different underlying mechanisms for the lack of improvement in the second training block (i.e., *n*-back with spoken words) for these two cases: E.Q. seemed to reach asymptote by the fourth session in the second training block and change in her performance may have gone undetected due to statistical properties of the Spearman correlation coefficient (i.e., it measures linear relationships) used to test for performance improvements. In case of M.N., however, results of the motivation questionnaires suggest that the lack of improvement may be due to decreasing motivation and engagement with the training activity and/or to the fact that an *n*-back task including spoken stimuli was more difficult for her than another including pictures. In sum, participants improved in the training tasks, and more importantly, the amount of improvement was comparable to that observed in healthy young adults in similar tasks (Novick et al., 2014).

Consistent with our previous study (Zakariás et al., 2016), we detected a mixed pattern of transfer. With respect to far transfer, all participants improved at least in three outcome measures out of the five. These included measures of spoken sentence comprehension (i.e., TROG-D, Sätze verstehen, Token test), functional communication (ANELT), and everyday memory (Everyday memory questionnaire). Crucially, follow-up results suggest that improvements in spoken sentence comprehension were also maintained at six weeks after completion of the study for two participants. Although psychometric properties are not known for all the far transfer tasks we used, results of a previous study indicates that the TROG has high test-retest reliability (*r* = .99 in a group of five people with different types and severity of aphasia, see Zakariás et al., 2016). Furthermore, the two parallel versions of the ANELT correlate with each other to a very high degree (Blomert et al., 1994). Both the TROG and ANELT could be used to evaluate treatment effects in spoken sentence comprehension and functional communication respectively. The current results are in line with previous findings of Eom and Sung (2016) and Zakariás et al. (2016), who also found improvement after a WM training on spoken sentence comprehension. To our knowledge, this is the first study showing transfer effects after WM training on functional communication in aphasia.

With regards to the specificity of transfer effects on spoken sentence comprehension, we detected improvements on: (1) non-canonical number marked (object-verb-subject) sentences, (2) non-canonical sentences including varying syntactic structures, such as case marked and number marked object-verb-subject sentences and right-branching and centre-embedding object relative clauses, and (3) canonical sentences including case marked and number marked subject-verb-object sentences and right-branching and centre-embedding subject relative clauses in some individuals. What mechanisms can account for these improvements? A number of studies have suggested that WM supports parsing and interpretation (i.e., construction of the syntactic structure of a sentence and the use of this structure to determine sentence meaning, respectively) and is majorly involved in processing syntactically complex sentences, such as object-relative clauses (see Just and Carpenter, 1992 for review; Haarmann et al., 1997). Just and Carpenter (1992) argued that the same pool of WM resources tapped by WM tasks is also used in sentence processing. By contrast, Caplan and colleagues (2013) proposed that memory mechanisms captured by traditional WM tasks (e.g., WM span and *n*-back) do not support the on-line, automatic processing of syntactic information, but are engaged in a later stage of sentence comprehension, namely the revision of the previously encountered, inaccurately interpreted information, and the use of the product of the comprehension to perform a task (e.g., in a picture-matching task keeping sentence meaning in mind while analysing and interpreting the visual scenes and comparing them to the meaning of the sentence). This is called post-interpretive or expanded comprehension (Caplan et al., 2013). Our results showing improvements on both canonical and non-canonical structures after WM training in IWA suggest that the use of WM in sentence processing is less specific to syntactic structures but may play a role in more general processes involved in the later stage of sentence comprehension (post-interpretive comprehension). This aspect is particularly important in everyday tasks that involve sentence comprehension (e.g., extracting meaning from conversations, understanding information from the news).

With respect to near transfer, the pattern of improvements in the WM tasks suggests that very near transfer occurred. Gains detected in the *n*-back with letters but not in the running span suggest that the improvements were task specific rather than process specific.

What linguistic and cognitive profiles make participants likely benefit from WM training? According to Caplan et al. (2013) and Fedorenko (2014), WM provides extra computational resources or alternative routes for resolving the possible problems encountered during language comprehension. These theories also imply that WM training can be most beneficial for IWA demonstrating substantial WM deficits, because potential improvements on WM allow them to utilize extra resources during language comprehension. To investigate the potential relationship between initial WM, language comprehension abilities, and improvement on spoken sentence comprehension after training, we performed a Spearman rank correlation on the data of the current study and the data collected in our previous study (Zakariás et al., 2016). This comparison was possible, because some of the WM tasks and the spoken sentence comprehension tests used in the two studies were identical in terms of task design and procedures (i.e., *n*-back with letters), or were standardized in both languages (i.e., TROG). Results of the analysis revealed a relationship between initial spoken sentence comprehension ability and training outcome (*rho* = -.754, p = .084), suggesting that the more severe the spoken sentence comprehension deficit was at the beginning of training, the more it improved after training. However, we could not find any relationship between initial WM and improvement on spoken sentence comprehension after training. The lack of a significant correlation between these variables could be due to the small number of data entered into the analyses.

Results of Zakariás et al. (2016) and the present study also suggest that the extent of improvement in an *n*-back training task is not necessarily proportional to the improvement in the transfer tasks in aphasia, as proposed by others investigating transfer in other populations, such as children and healthy young adults (e.g., Jaeggi et al., 2014; Waris, Soveri, & Laine, 2015, respectively). This may be related to the fact that in aphasia the extent of improvement on the trained processes and a complex interaction of intact and impaired functions affect training outcome and benefit to untrained functions.

The lack of significant improvement in everyday memory can be explained by the fact that participants had only mild impairments already at the beginning of the training, and therefore, there was not enough room for improvement. However, it is still difficult to interpret the negative change in performance for M.N. One possible explanation could be that M.N. and her daughter (who rated the everyday memory questionnaire) did not have everyday contact and communication during the time of study (i.e., they did not live together). Insufficient communication or biases might have lead to false estimation (i.e., in this case overestimation) of the memory failures.

4.1. Limitations of the present study

There are a few limitations of the current study that could inform future research on WM training in aphasia. Because we assessed experimental outcome measures using a multiple baseline design, we did not include a control group. For a few conditions, however, baseline variability was too large to provide stable baseline estimates, which could have led to some outcome effects overestimated or going undetected. For these conditions therefore, both significant and non-significant effects should be interpreted with caution. In particular, experimental tasks used for multiple baseline assessments can benefit from more baselines.

A further concern relates to the Sätze verstehen test to assess specificity of transfer effects. The lack of significant effects in most conditions in this measure (despite significant effects on aggregated scores) could be a result of low statistical power due to only a small number of sentences per condition. Future research with a larger number of examples of each sentence type might allow for a better understanding of the underlying mechanisms of transfer effects on sentence comprehension.

Although the single-case experimental design employed ensures valid estimation of effects at the individual level, the large individual differences call for future research to clarify how far these results are generalizable to population level of IWA. A more feasible goal for prospective studies would be to identify a few mechanisms that may generalize to at least a subpopulation of IWA. To achieve this goal, more detailed analyses of individual differences on larger but yet homogenous samples may be required.

In conclusion, the present study is the first systematic investigation of transfer effects of training higher-level WM functions on functional communication and everyday functions in aphasia. Our results suggest that WM can improve through intensive computerized training in chronic aphasia and this improvement can lead to improvements in spoken sentence comprehension and functional communication.

Acknowledgements

We are grateful to E.Q., I.B., and M.N. for their participation, and thank Leonie Lampe for her help with preparing the assessment and training materials, Helene Gerda Killmer, Susanne Pelke, Thomas Hinsdorf, and Vanessa Dawn Löffler for their help with the data collection, and Jan Ries for creating the figures.

References

Baddeley, A. (2012). Working memory: theories, models, and controversies. *Annual review of Psychology, 63,* 1-29. https://doi.org/10.1146/annurev-psych-120710-100422

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, *8,* 47-89. https://doi.org/10.1016/S0079-7421(08)60452-1

Barker-Collo, S., Feigin, V. L., Parag, V., Lawes, C. M. M., & Senior, H. (2010). Auckland stroke outcomes study part 2: Cognition and functional outcomes 5 years poststroke. *Neurology, 75*(18), 1608-1616. http://dx.doi.org/10.1212/WNL.0b013e3181fb44c8

Blomert, L., Kean, M. L., Koster, C., & Schokker, J. (1994). Amsterdam—Nijmegen everyday language test: construction, reliability and validity. *Aphasiology*, *8*(4), 381–407. https://doi.org/10.1080/02687039408248666

Brownsett, S. L., Warren, J. E., Geranmayeh, F., Woodhead, Z., Leech, R., & Wise, R. J. (2013). Cognitive control and its impact on recovery from aphasic stroke. *Brain, 137*(1), 242-254. https://doi.org/10.1093/brain/awt289

Brunner, C., & Steiner, J. (1994). ANELT Amsterdam-Nijmegen Everyday Language Test – Deutsche Fassung. *Entwickelt von Blomert & Buslach.*

Burchert, F., Lorenz, A., Schröder, A., De Bleser, R., & Stadie, N. (2011). *Sätze verstehen. Neurolinguistische Materialien für die Untersuchung von syntaktischen Störungen beim Satzverständnis.* Hofheim: NAT-Verlag.

Caplan, D., Michaud, J., & Hufford, R. (2013). Short-term memory, working memory, and syntactic comprehension in aphasia. *Cognitive Neuropsychology, 30*(2), 77-109. http://dx.doi.org/10.1080/02643294.2013.803958

Caplan, D., & Waters, G. (2013). Memory mechanisms supporting syntactic comprehension. *Psychonomic Bulletin & Review, 20*(2), 243-268. https://doi.org/10.3758/s13423-012-0369-9

Caspari, I., Parkinson, S. R., LaPointe, L. L., & Katz, R. C. (1998). Working memory and aphasia. *Brain and Cognition*, *37*(2), 205–223. https://doi.org/10.1006/brcg.1997.0970

Christensen, S. C., & Wright, H. H. (2010). Verbal and non-verbal working memory in aphasia: What three n-back tasks reveal. *Aphasiology, 24*(6-8), 752-762. http://dx.doi.org/10.1080/02687030903437690

Coelho, C. (2005). Direct attention training as a treatment for reading impairment in mild aphasia. *Aphasiology*, *19*(3–5), 275–283. https://doi.org/10.1080/02687030444000741

Cohen, J. D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., & Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature 386*(6625), 604-608. http://hdl.handle.net/2027.42/62513

Collette, F., Van der Linden, M., Laureys, S., Arigoni, F., Delfiore, G., Degueldre, C., ... & Salmon, E. (2007). Mapping the updating process: common and specific brain activations across different versions of the running span task. *Cortex, 43*(1), 146-158. https://doi.org/10.1016/S0010-9452(08)70452-0

Cowan, N. (2008). What are the differences between long-term, short-term, and working memory? *Progress in Brain Research, 169*, 323-338. https://doi.org/10.1016/S0079-6123(07)00020-9

DeDe, G., Ricca, M., Knilans, J., & Trubl, B. (2014). Construct validity and reliability of working memory tasks for people with aphasia. *Aphasiology*, *28*(6), 692–712. https://doi.org/10.1080/02687038.2014.895973

Duffin, J. T., Collins, D. R., Coughlan, T., O'Neill, D., Roche, R. A., & Commins, S. (2012). Subtle memory and attentional deficits revealed in an Irish stroke patient sample using domain-specific cognitive tasks. *Journal of Clinical and Experimental Neuropsychology, 34*(8), 864-875. http://dx.doi.org/10.1080/13803395.2012.690368

Engle, R. W. (2002). Working Memory Capacity as Executive Attention. *Current Directions in Psychological Science*, *11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

Eom, B., & Sung, J. E. (2016). The Effects of Sentence Repetition–Based Working Memory Treatment on Sentence Comprehension Abilities in Individuals With Aphasia. *American Journal of Speech-Language Pathology*, *25*(4S), S823–S838. https://doi.org/10.1044/2016_AJSLP-15-0151

Fedorenko, E. (2014). The role of domain-general cognitive control in language comprehension. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00335

Fox-Boyer, A. V. (2013). *TROG-D: Test zur Überprüfung des Grammatikverständnisses.* Schulz-Kirchner Verlag GmbH.

Francis, D., Clark, N., & Humphreys, G. (2003). The treatment of an auditory working memory deficit and the implications for sentence comprehension abilities in mild “receptive” aphasia. *Aphasiology*, *17*(8), 723–750. https://doi.org/10.1080/02687030344000201

Frankel, T., Penn, C., & Ormond- Brown, D. (2007). Executive dysfunction as an explanatory basis for conversation symptoms of aphasia: A pilot study. *Aphasiology*, *21*(6–8), 814–828. https://doi.org/10.1080/02687030701192448

Franklin, R. D., Allison, D. B., & Gorman, B. S. (2014). *Design and analysis of single-case research*. Psychology Press. Retrieved from https://books.google.de/books?hl=en&lr=&id=aKSYAgAAQBAJ&oi=fnd&pg=PP1&dq=Design+and+Analysis+of+Single-Case+Research+edited+by+Ronald+D.+Franklin,+David+B.+Allison,+Bernard+S.+Gorman&ots=sDycYBstdB&sig=nwGyoJNf_LW26x7GGHdnjHmM4Hc

Fridriksson, J., Nettles, C., Davis, M., Morrow, L., & Montgomery, A. (2006). Functional communication and executive function in aphasia. *Clinical*

Linguistics & Phonetics, *20*(6), 401–410. https://doi.org/10.1080/02699200500075781

Friedman, N. P., & Miyake, A. (2004). The relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133*(1), 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Friedmann, N., & Gvion, A. (2003). Sentence comprehension and working memory limitation in aphasia: A dissociation between semantic-syntactic and phonological reactivation. *Brain and Language*, *86*(1), 23–39. https://doi.org/10.1016/S0093-934X(02)00530-8

Green, D. W., Grogan, A., Crinion, J., Ali, N., Sutton, C., & Price, C. J. (2010). Language control and parallel recovery of language in individuals with aphasia. *Aphasiology*, *24*(2), 188–209. https://doi.org/10.1080/02687030902958316

Haarmann, H. J., Just, M. A., & Carpenter, P. A. (1997). Aphasic sentence comprehension as a resource deficit: A computational approach. *Brain and Language, 59*(1), 76-120. https://doi.org/10.1006/brln.1997.1814

Harris, L., Olson, A., & Humphreys, G. (2014). The link between STM and sentence comprehension: A neuropsychological rehabilitation study. *Neuropsychological Rehabilitation*, *24*(5), 678–720. https://doi.org/10.1080/09602011.2014.892885

Helm-Estabrooks, N., & Albert, M. L. (1991). *Manual of aphasia therapy.* Pro Ed.

Huber, W. (1983). *Aachener Aphasie-Test:(AAT).* Hogrefe, Verlag für Psychologie.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National*

Academy of Sciences, *105*(19), 6829–6833. https://doi.org/10.1073/pnas.0801268105

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086. https://doi.org/10.1073/pnas.1103228108

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition*, *42*(3), 464–480. https://doi.org/10.3758/s13421-013-0364-z

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149. http://dx.doi.org/10.1037/0033-295X.99.1.122

Kane, M. J., Conway, A. R., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: a question of construct validity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(3), 615. http://dx.doi.org/10.1037/0278-7393.33.3.615

Katz, B., Jaeggi, S., Buschkuehl, M., Stegman, A., & Shah, P. (2014). Differential effect of motivational features on training improvements in school-based cognitive training. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00242

Kay, J., Lesser, R., & Coltheart, M. (1996). Psycholinguistic assessments of language processing in aphasia (PALPA): An introduction. *Aphasiology, 10*(2), 159-180. http://dx.doi.org/10.1080/02687039608248403

Keil, K., & Kaszniak, A. W. (2002). Examining executive function in individuals with brain injury: A review. *Aphasiology*, *16*(3), 305–335. https://doi.org/10.1080/02687030143000654

Lambon Ralph, M. A., Snell, C., Fillingham, J. K., Conroy, P., & Sage, K. (2010). Predicting the outcome of anomia therapy for people with aphasia post CVA: Both language and cognitive status are key predictors. *Neuropsychological Rehabilitation*, *20*(2), 289–305. https://doi.org/10.1080/09602010903237875

Lind, M., Kristoffersen, K. E., Moen, I., & Simonsen, H. G. (2009). Semi-spontaneous oral text production: Measurements in clinical practice. *Clinical Linguistics & Phonetics*, *23*(12), 872–886. https://doi.org/10.3109/02699200903040051

Lindeløv, J. K., Dall, J. O., Kristensen, C. D., Aagesen, M. H., Olsen, S. A., Snuggerud, T. R., & Sikorska, A. (2016). Training and transfer effects of N-back training for brain-injured and healthy subjects. *Neuropsychological Rehabilitation*, *26*(5–6), 895–909. https://doi.org/10.1080/09602011.2016.1141692

Lorenz, A., Heide, J., & Burchert, F. (2014). Compound naming in aphasia: Effects of complexity, part of speech, and semantic transparency. *Language, Cognition, and Neuroscience, 29,* 88-106. doi:10.1080/01690965.2013.766357

Luna, C. Z. (2011). Functional communication in chronic aphasia and executive function: The effect of treating cognitive flexibility [Clinical Aphasiology Conference Paper]. Retrieved April 18, 2017, from http://aphasiology.pitt.edu/2243/

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: Cross-Linguistic Easy-Access Resource for Phonological and Orthographic Neighborhood Densities. *PLOS ONE*, *7*(8), e43230. https://doi.org/10.1371/journal.pone.0043230

Martin, N., Kohen, F., Kalinyak-Fliszar, M., Soveri, A., & Laine, M. (2012). Effects of working memory load on processing of sounds and meanings of words in aphasia. *Aphasiology*, *26*(3–4), 462–493. https://doi.org/10.1080/02687038.2011.619516

Mayer, J. F., & Murray, L. L. (2002). Approaches to the treatment of alexia in chronic aphasia. *Aphasiology*, *16*(7), 727–743. https://doi.org/10.1080/02687030143000870

Mayer, J. F., & Murray, L. L. (2012). Measuring working memory deficits in aphasia. *Journal of Communication Disorders, 45*(5), 325-339. https://doi.org/10.1016/j.jcomdis.2012.06.002

Mayer, J. F., Mitchinson, S. I., & Murray, L. L. (2016). Addressing concomitant executive dysfunction and aphasia: Previous approaches and the new brain budget protocol. *Aphasiology*, *0*(0), 1–24. https://doi.org/10.1080/02687038.2016.1249333

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48–58. https://doi.org/10.1080/02701367.1989.10607413

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex “frontal lobe” tasks: A latent variable analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Morgan, D. L., & Morgan, R. K. (2008). *Single-case research methods for the behavioral and health sciences*. SAGE Publications.

Murray, L. L., Keeton, R. J., & Karcher, L. (2006). Treating attention in mild aphasia: Evaluation of attention process training-II. *Journal of Communication Disorders, 39*(1), 37-61. https://doi.org/10.1016/j.jcomdis.2005.06.001

Nicholas, L. E., & Brookshire, R. H. (1993). A System for quantifying the informativeness and efficiency of the connected speech of adults with aphasia. *Journal of Speech, Language, and Hearing Research*, *36*(2), 338–350. https://doi.org/10.1044/jshr.3602.338

Nicholas, M., Hunsaker, E., & Guarino, A. J. (2017). The relation between language, non-verbal cognition and quality of life in people with aphasia. *Aphasiology*, *31*(6), 688–702. https://doi.org/10.1080/02687038.2015.1076927

Nickels, L., Howard, D., & Best, W. (1997). Fractionating the articulatory loop: Dissociations and associations in phonological recoding in aphasia. *Brain and Language*, *56*(2), 161–182. https://doi.org/10.1006/brln.1997.1732

Novick, J. M., Hussey, E., Teubner-Rhodes, S., Harbison, J. I., & Bunting, M. F. (2014). Clearing the garden-path: improving sentence processing through cognitive control training. *Language, Cognition and Neuroscience*, *29*(2), 186–217. https://doi.org/10.1080/01690965.2012.758297

Novick, J. M., Kan, I. P., Trueswell, J. C., & Thompson-Schill, S. L. (2009). A case for conflict across multiple domains: Memory and language impairments following damage to ventrolateral prefrontal cortex. *Cognitive Neuropsychology*, *26*(6), 527–567. https://doi.org/10.1080/02643290903519367

Paek, E. J, & Murray, L. L. (2015). Working memory approach for language and memory deficits in an individual with chronic aphasia: a case study. *Ehearsay:*

Electronic Journal Of The Ohio Speech- language Hearing Association, 5(1), 86-98.

Peach, R. K., Nathan, M. R., & Beck, K. M. (2017). Language-specific attention treatment for aphasia: Description and preliminary findings. *Seminars in Speech and Language*, *38*(1), 005–016. https://doi.org/10.1055/s-0036-1597260

Penn, C., Frankel, T., Watermeyer, J., & Russell, N. (2010). Executive function and conversational strategies in bilingual aphasia. *Aphasiology*, *24*(2), 288–308. https://doi.org/10.1080/02687030902958399

Pollack, I., Johnson, L. B., & Knaff, P. R. (1959). Running memory span. *Journal of experimental Psychology, 57*(3), 137-146. http://dx.doi.org/10.1037/h0046137

Purdy, M. (2002). Executive function ability in persons with aphasia. *Aphasiology*, *16*(4–6), 549–557. https://doi.org/10.1080/02687030244000176

Ramsberger, G. (2005). Achieving conversational success in aphasia by focusing on non-linguistic cognitive skills: A potentially promising new approach. *Aphasiology*, *19*(10–11), 1066–1073. https://doi.org/10.1080/02687030544000254

Robinson, G., Blair, J., & Cipolotti, L. (1998). Dynamic aphasia: an inability to select between competing verbal responses? *Brain*, *121*(1), 77–89. https://doi.org/10.1093/brain/121.1.77

Rossion, B., & Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. *Perception*, *33*(2), 217–236. https://doi.org/10.1068/p5117

Salis, C. (2012). Short-term memory treatment: Patterns of learning and generalisation to sentence comprehension in a person with aphasia.

Neuropsychological Rehabilitation, 22(3), 428-448. http://dx.doi.org/10.1080/09602011.2012.656460

Salis, C., Hwang, F., Howard, D., & Lallini, N. (2017). Short-term and working memory treatments for improving sentence comprehension in aphasia: A review and a replication study. *Seminars in Speech and Language*, *38*(1), 029–039. https://doi.org/10.1055/s-0036-1597262

Sharp, D. J., Turkheimer, F. E., Bose, S. K., Scott, S. K., & Wise, R. J. S. (2010). Increased frontoparietal integration after stroke and cognitive recovery. *Annals of Neurology*, *68*(5), 753–756. https://doi.org/10.1002/ana.21866

Sinotte, M. P., & Coelho, C. A. (2007). Attention training for reading impairment in mild aphasia: A follow-up study. *NeuroRehabilitation*, *22*(4), 303–310.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174–215. https://doi.org/10.1037/0278-7393.6.2.174

Stadie, N., Cholewa, J., & De Bleser, R. (2013). *LEMO 2.0: Lexikon modellorientiert: Diagnostik für Aphasie, Dyslexie und Dysgraphie.* NAT-Verlag.

Stewart, F. M., Sunderland, A., & Sluman, S. M. (1996). The nature and prevalence of memory disorder late after stroke. *British Journal of Clinical Psychology*, *35*(3), 369–379. https://doi.org/10.1111/j.2044-8260.1996.tb01191.x

Sunderland, A., Harris, J. E., & Baddeley, A. D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 341–357. https://doi.org/10.1016/S0022-5371(83)90229-3

Sung, J. E., McNeil, M. R., Pratt, S. R., Dickey, M. W., Hula, W. D., Szuminsky, N. J., & Doyle, P. J. (2009). Verbal working memory and its relationship to sentence- level reading and listening comprehension in persons with aphasia. *Aphasiology*, *23*(7–8), 1040–1052. https://doi.org/10.1080/02687030802592884

Tompkins, C. A., Bloise, C. G. R., Timko, M. L., & Baumgaertner, A. (1994). Working memory and inference revision in brain-damaged and normally aging adults. *Journal of Speech, Language, and Hearing Research*, *37*(4), 896–912. https://doi.org/10.1044/jshr.3704.896

Vallat-Azouvi, C., Pradat-Diehl, P., & Azouvi, P. (2014). Modularity in rehabilitation of working memory: A single-case study. *Neuropsychological Rehabilitation, 24*(2), 220-237. http://dx.doi.org/10.1080/09602011.2014.881294

Vallat, C., Azouvi, P., Hardisson, H., Meffert, R., Tessier, C., & Pradat-Diehl, P. (2005). Rehabilitation of verbal working memory after left hemisphere stroke. *Brain Injury, 19*(13), 1157-1164. http://dx.doi.org/10.1080/02699050500110595

Waris, O., Soveri, A., & Laine, M. (2015). Transfer after working memory updating training. *PLOS ONE*, *10*(9), e0138734. https://doi.org/10.1371/journal.pone.0138734

Wilson, B., Cockburn, J., Baddeley, A., & Hiorns, R. (1989). The development and validation of a test battery for detecting and monitoring everyday memory problems. *Journal of Clinical and Experimental Neuropsychology*, *11*(6), 855–870. https://doi.org/10.1080/01688638908400940

Wright, H. H., Downey, R. A., Gravier, M., Love, T., & Shapiro, L. P. (2007). Processing distinct linguistic information types in working memory in aphasia. *Aphasiology*, *21*(6–8), 802–813. https://doi.org/10.1080/02687030701192414

Zakariás, L., Keresztes, A., Demeter, G., & Lukács, Á. (2013). A specific pattern of executive dysfunctions in transcortical motor aphasia. *Aphasiology*, *27*(12), 1426–1439. https://doi.org/10.1080/02687038.2013.835783

Zakariás, L., Keresztes, A., Marton, K., & Wartenburger, I. (2016). Positive effects of a computerised working memory and executive function training on sentence comprehension in aphasia. *Neuropsychological Rehabilitation*, 1–18. http://dx.doi.org/10.1080/09602011.2016.1159579

Table 1. *Summary of WM treatments including outcome measures of spoken sentence comprehension, spoken discourse, and verbal communication in individuals with aphasia*

Studies	*Participant(s)*	*Treatment procedures*	*Outcomes on language*
Francis et al. (2003)	*n* = 1 (mild chronic aphasia)	Sentence repetition	− TROG, TT, and active reversible sentences
Harris et al. (2013)	*n* = 2 (Broca's aphasia [DS], mild aphasia [AK])	Repetition and recognition tasks with words and non-words	⇑ for DS in semantically anomalous sentence judgements and sentence-picture matching (PALPA 55)
Salis (2012)	*n* = 1 (severe TMA)	Matching listening span with nouns	⇑ TROG; − TT
Paek and Murray (2015)	*n* = 1 (mild anomic aphasia)	*N*-back with pictures/written words, updating with pictures/written words, reading span involving grammaticality judgments and category naming, naming with spaced retrieval, opposite sentence training, reconstitution of words from oral spelling	− RTT; ⇑ %CIUs, CIUs/min in story-telling
Zakariás et al. (2016)	*n* = 3 (moderate chronic anomic [KK] and TMA [BL, BB])	Adaptive *n*-back with letters	⇑ for KK and BL in the TROG-H
Eom and Sung (2016)	*n* = 6 (Broca's, anomic, and Wernicke aphasia)	Repetition-based treatment protocol (active sentences with two- and three-argument verbs, passive sentences, conjoined sentences, and centre-embedded sentences with a subject-relative clause)	⇑ for five participants in sentence picture matching (Sung, 2015) including active sentences with two-argument verbs, active sentences with three-argument verbs, and passive counterparts of active sentences with two-argument verbs
Salis et al. (2017)	*n* = 5 (moderate chronic aphasia)	Matching listening span with nouns	− TROG, TT, ⇑ in the CETI for one participant

Note. ⇑: improvement in the task; −: no change in the task; TROG: Test for the Reception of Grammar; TT: Token test; TMA: transcortical motor aphasia; %CIUs: percent of correct information units; CIUs/min: correct information units per minute; RTT: Revised Token test; TROG-H: Hungarian version of the Test for the Reception of Grammar; CETI: Communication Effectiveness Index; PALPLA: Psycholinguistic assessments of language processing in aphasia.

Table 2. *Background description of the participants*

	E.Q.	I.B.	M.N.
Gender	F	F	F
Age (years)	39	77	51
Education (years)	10	12	10
Etiology	CVA	CVA	CVA
Lesion	Infarct of the left MCA	Infarct of the left MCA	Infarct of the left MCA
Time post onset (years)	6	25	15
Aphasia type (AAT profile)	99.3% Broca's, 0.7% anomic	Unclassified	Unclassified
AAT (%)			
Token	60	56	30
Repetition	79.3	72.6	69.3
Written language	90	72.2	36.6
Naming	85.83	80.83	56.6
Comprehension	91.66	88.33	70.83
Spoken words	100	93.3	66.66
Spoken sentences	80	80	73.33
Written words	100	90	73.33
Written sentences	86.66	90	70
TROG-D (%)	77.38	76.19	53.57
Listening span – verbal WM (span)	2	2	1
Corsi block tapping – visuo-spatial WM (span)	5	4	5

Note. CVA: cerebrovascular accident; MCA: middle cerebral artery; AAT: German version of the Aachen Aphasia Test; TROG-D: German version of the Test for the Reception of Grammar; WM: working memory; note that AAT scores were obtained one and two years before the present study (for M.N., and for E.Q. and I.B., respectively). Other assessment data was obtained at the beginning of the study.

Table 3. *Improvements on the outcome measures*

	Case							
	E.Q.		I.B.		M.N.		Group	
Outcome measure	Pre-post	FU	Pre-post	FU	Pre-post	FU	Pre-post	FU
N-back with letters								
1-back								
2-back	⇑		⇑					
3-back	⇑		⇑					
Running span								
TROG-D					⇑	✓		⇗
Token test	⇗				⇗		⇗	
Sätze verstehen								
Short irreversible								
Long irreversible								
Case-marked SVO								
Case-marked OVS								
Number-marked SVO								
Number-marked OVS	⇑							
Right-branching SRC							⇗	
Right-branching ORC								
Centre-embedded SRC								
Centre-embedded ORC							⇗	
Total			⇗				⇗	
Canonical			⇑					
Non-canonical	⇑							
ANELT								
Understandability			⇑					
Number of words					⇑			
CIUs					⇑		⇗	
%CIUs			⇑				⇗	
CIUs/min			⇗				⇗	
Words/min								
EMQ								
Speech	⇗		⇗		⇓			
Learning new things					⇗			

Note. FU: follow-up; TROG-D: German version of the Test for the Reception of Grammar; SVO: subject-verb-object; OVS: object-verb-subject; SRC: subject relative clauses; ORC: object-relative clauses; ANELT: Amsterdam-Nijmegen Everyday Language Test; CIUs: correct information units; %CIUs: percent of correct information units; CIUs/min: correct information units per minute; EMQ: Everyday memory questionnaire; ⇑ and ⇗ indicate a statistically significant improvement and a tendency for improvement, respectively; ⇓ indicates a statistically significant decrease; empty grey cells indicate that data was available, but did not produce statistically significant change; ✓ shows maintenance of performance at follow-up. Note that performance was close to ceiling already at the beginning of the training in the letter 1-back, EMQ 'Speech', and 'Short irreversible', 'Long irreversible', 'Case-marked SVO', 'Number-marked SVO' for E.Q. and I.B., and 'Long irreversible' for M.N. For the raw date obtained in the outcome measures, see the supplementary material – Table S4.

Figure captions

Figure 1. Design and tasks used in the study. Participants were randomly assigned to the order of the training blocks. Initial assessment was used to assess suitability in the present study. Pretest and posttest took 2.5 weeks each. Training blocks took 2-3 weeks each. The study lasted altogether ~10 weeks (30 sessions). Follow-up was conducted 4-6 weeks after completion of the posttest.

Figure 2. Two *n*-back tasks (pictures, spoken words) with “lures” used as training tasks, illustrated here with three levels of difficulty comprising three lure levels within the 2-back level. Participants had to perform three lure levels before *n* increased. Level 3: 2-back with no lures. Level 4: 2-back with lures at $n+1$ position. Level 5: 2-back with lures at $n+1$ and $n-1$ position. Note that at the 1-back level there could be no lures at the $n-1$ position, hence there are only two difficulty levels before level 3: 1-back with no lures, and 1 back with lures at the $n+1$ position.

Figure 3. Performance on the training tasks during the 16 sessions of training. I.B. improved significantly ($p < .05$) across sessions in both training tasks, whereas E.Q. and M.N. improved statistically significantly only in the first training comprising the *n*-back with pictures ($p < .05$).

Figure 4. Mean scores of interest/enjoyment, perceived competence, and effort/importance over the sixteen sessions of training.

Supplementary Material

Table S1. *Questions of the everyday memory questionnaire in German and English*

Speech		
1	Er/Sie bringt Namen von Freunden oder Verwandten durcheinander oder nennt sie bei einem falschen Namen.	He/She confuses the names of friends or relatives or calls them by the wrong names.
2	Er/Sie bringt Namen von geläufigen Dingen durcheinander oder nennt sie bei einem falschen Namen.	He/She confuses the names of common things or uses the wrong names.
3	Ihm/Ihr liegen Wörter auf der Zunge. Er/Sie kennt das Wort aber kann es nicht finden.	He/She has words on the tip of his/her tongue. He/She knows what it is but can't quite find it.
4	Er/Sie vergisst Dinge, die einige Minuten zuvor gesagt wurden. Zum Beispiel etwas, was der Ehepartner oder ein Freund gerade gesagt hat.	He/She forgets something that he/she was told a few minutes earlier; for instance, something his/her spouse or a friend has just said.
5	Er/Sie vergisst, was ihm/ihr gestern oder vor einigen Tagen erzählt wurde.	He/She forgets something he/she was told yesterday or a few days earlier.
6	Er/Sie wiederholt Dinge, die er/sie kurz zuvor gesagt hat oder stellt die gleichen Fragen mehrmals.	He/She repeats something he has just said or asks the same question several times.
7	Er/Sie vergisst, was er/sie gerade gesagt hat. Dabei sagt er/sie möglicherweise etwas wie "Worüber habe ich gerade gesprochen?"	He/She forgets what he has just said. Thereby, he possibly says something like "What have I just been talking about?"
8	Er/Sie ist nicht in der Lage, dem zu folgen, was jemand erzählt. In einem Gespräch verliert er/sie den Faden.	He/She loses track of what someone tells him/her. During a conversation, he loses the thread.
9	Er/Sie beginnt etwas zu sagen, vergisst dann aber, worüber er/sie eigentlich sprechen wollte.	He/She starts to say something, but then forgets what he actually wanted to talk about.
10	Er/Sie schweift ab und spricht über unwichtige und irrelevante Dinge.	He/She gets off the point and speaks about unimportant or irrelevant things.
11	Er/Sie vergisst, anderen wichtige Dinge zu erzählen. Zum Beispiel vergisst er/sie, eine Nachricht weiterzuleiten oder jemanden an etwas zu erinnern.	He/She forgets to tell others something important. For instance, he forgets to pass on a message or to remind someone of something.
12	Er/Sie bringt Details von dem durcheinander, was ihm/ihr jemand erzählt hat.	He/She mixes up the details of what someone has told him.
13	Er/Sie wiederholt Geschichten oder Witze, die er/sie bereits erzählt hat.	He/She repeats a story or joke he has said before.
Faces and places		
14	Er/Sie vergisst, wo er/sie Dinge hingelegt hat. Er/Sie verlegt Dinge im Haus.	He/She forgets where he put something. He misplaces things around the house.
15	Er/Sie erkennt Angehörige und Freunde nicht.	He/She does not recognise relatives and friends.
16	Er/Sie erkennt Fernsehcharaktere oder andere Berühmtheiten nicht.	He/She does not recognise television characters or other famous people.

17	Er/Sie verläuft sich oder geht auf einem Weg oder Spaziergang in die falsche Richtung, den er/sie schon oft gegangen ist.	He/She gets lost or takes the wrong direction on a route or walk that he went on often.
18	Er/Sie erkennt Orte nicht, von denen ihm/ihr gesagt wurde, dass er/sie dort schon oft gewesen sei.	He/She does not recognise places he was told that he has often been to before.
19	Es fällt ihm/ihr schwer, im Fernsehen der Handlung zu folgen.	It is hard for him/her to follow the storyline when watching TV.
Actions		
20	Er/Sie vergisst regelmäßige Handlungen, die er/sie sonst ein- oder zweimal am Tag durchführen würde.	He/She forgets regular activities that he would normally do once or twice a day.
21	Er/Sie stellt fest, dass er/sie eine regelmäßige Handlung ausversehen zweimal durchgeführt hat.	He/She discovers that he did some regular activity twice by mistake.
22	Er/Sie muss überprüfen, ob er/sie alles getan hat, was er/sie tun sollte.	He/She has to check whether he has done everything he ought to.
23	Er/Sie vergisst, was er/sie gestern gemacht hat oder bringt die Details von dem durcheinander, was passiert ist.	He/She forgets what he did yesterday or getting the details of what happened mixed up and confused.
24	Er/Sie fängt an, Dinge zu tun und vergisst aber währenddessen, was er/sie eigentlich tun wollte. Dabei sagt er/sie möglicherweise etwas wie "Was tue ich hier?"	He/She starts doing something, but then forgets what he was intending to do. Thereby, he possibly says something like "What am I doing here?"
25	Er/Sie ist geistesabwesend. Er/Sie tut Dinge, die er/sie nicht wirklich vorhatte.	He/She is absentminded. He does things that he did not really intend to do.
Learning new things		
26	Er/Sie erinnert sich nicht an den Namen von jemandem, den er/sie vor kurzem zum ersten Mal getroffen hat.	He/She is not able to remember the name of someone he met for the first time recently.
27	Er/Sie erkennt Menschen nicht, die er/sie vor kurzem zum ersten Mal getroffen hat.	He/She does not recognise people he met for the first time recently.
28	Er/Sie verläuft sich auf einem Weg oder Spaziergang, den er/sie vorher nur ein- oder zweimal gegangen ist.	He/She gets lost on a route or walk that he has only gone on once or twice before.
29	Es gelingt ihm/ihr nicht, eine neue Fertigkeit, wie z.B. ein Spiel oder den Umgang mit einem Gerät, zu erlernen, wenn er/sie es ein- oder zweimal geübt hat.	He/She is not able to pick up a new skill, such as a game or handling a new gadget, if he practised it once or twice.
30	Er/Sie kann mit Veränderung im Tagesablauf nicht umgehen. Er/Sie verfolgt dann irrtümlicherweise weiterhin die alte Routine.	He/She is not able to cope with changes in his daily routine. He then mistakenly keeps following the former routine.
31	Er/Sie vergisst, sich an Verabredungen zu halten.	He/She forgets to stick to agreements.

Rating scales for questionnaire presentations	
Section "Speech"	
(4) In etwa 60% oder mehr Fällen pro Tag (3) In weniger als 60% der Fälle pro Tag (2) Etwa einmal am Tag (1) Ein- oder zweimal in der Woche (0) Seltener als einmal in der Woche oder nie	(4) About 60 % or more of the cases in a day (3) Less than 60 % of the cases in a day (2) About once each day (1) Once or twice in a week (0) Less than once a week
Sections "Faces and places" and "Actions"	
(4) Mehrere Male am Tag (3) Etwa einmal am Tag (2) Ein- oder zweimal in der Woche (1) Seltener als einmal in der Woche (0) Nie	(4) Several times in a day (3) About once each day (2) Once or twice in a week (1) Less than once a week (0) Never
Section "Learning new things"	
(4) Jedes Mal (3) Häufiger (2) Nur manchmal (1) Sehr selten (0) Nie	(4) On every occasion (3) On every other occasion (2) Only sometimes (1) Very rarely (0) Never

Reference:

Sunderland, A., Harris, J. E., & Baddeley, A. D. (1983). Do laboratory tests predict everyday memory? A neuropsychological study. *Journal of Verbal Learning and Verbal Behavior*, *22*(3), 341–357. https://doi.org/10.1016/S0022-5371(83)90229-3

Table S2. *Questions of the motivation questionnaire in German and English*

Interest/enjoyment		
1	Wie gut hat Ihnen die Aufgabe heute gefallen? (überhaupt nicht gut; sehr gut)	How much did you enjoy the activity today? (not at all; a lot)
2	Wie viel Spaß hat Ihnen die Aufgabe gemacht? (überhaupt keinen Spaß; sehr viel Spaß)	How much fun was the activity to do? (not at all; a lot of fun)
3	Wie aufregend/spannend war die Aufgabe heute? (überhaupt nicht spannend; sehr spannend)	How exciting was the activity today? (not exciting at all; very exciting)
4	Wie gerne würden Sie die Aufgabe weiter üben, wenn wir Zeit dafür hätten? (überhaupt nicht gerne; sehr gerne)	How happily would you further practice the task if we had time?" (not gladly at all; very gladly)
Perceived competence		
5	Wie gut waren Sie heute in dieser Aufgabe?	How good were you at this activity today? (not good at all; very good)

	(überhaupt nicht gut; sehr gut)	
6	Wie gut haben Sie heute in dieser Aufgabe abgeschnitten, im Vergleich zu anderen Tagen? (überhaupt nicht gut; sehr gut)	How well did you do at this activity today, compared to other days? (not well at all; very well)
7	Wie zufrieden sind Sie mit Ihrer Leistung heute? (überhaupt nicht zufrieden; sehr zufrieden)	How satisfied are you with your performance today? (not satisfied at all; very satisfied)
Effort/importance		
8	Wie sehr haben Sie sich heute angestrengt? (überhaupt nicht; sehr)	How much effort did you put into this today? (no effort at all; a lot of effort)
9	Wie viel Mühe haben Sie sich heute mit dieser Aufgabe gegeben? (überhaupt keine Mühe; sehr viel Mühe)	How hard did you try on this activity today? (not hard at all; very hard)
10	Wie wichtig war es Ihnen, gut in dieser Aufgabe zu sein? (überhaupt nicht wichtig; sehr wichtig)	How important was for you to do well on this task? (not important at all; very important)

Note. The questionnaire was developed based on Jaeggi et al. (2011) and McAuley et al. (1989).

References:

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Shah, P. (2011). Short- and long-term benefits of cognitive training. *Proceedings of the National Academy of Sciences*, *108*(25), 10081–10086. https://doi.org/10.1073/pnas.1103228108

McAuley, E., Duncan, T., & Tammen, V. V. (1989). Psychometric properties of the Intrinsic Motivation Inventory in a competitive sport setting: A confirmatory factor analysis. *Research Quarterly for Exercise and Sport*, *60*(1), 48–58. https://doi.org/10.1080/02701367.1989.10607413

Table S3. *Stimuli and procedure of the listening span task*

Level 1 Sets	**Level 2 Sets**	**Level 3 Sets**	**Level 4 Sets**	**Level 5 Sets**
Set 1 Fische schwimmen im **Wasser** (T)	**Set 1** Im Kino schaut man einen **Film** (T) Milch ist **rot** (F)	**Set 4** Zucker ist **süß** (T) Berlin liegt neben **Rom** (F) Menschen essen Frühstück am **Abend** (F)	**Set 7** Giraffen haben einen langen **Hals** (T) Stühle können **essen** (F) Ein Fahrrad ist langsamer als ein **Bus** (T) Auf Konzerten gibt's **Musik** (T)	**Set 10** Hamster können **reden** (F) Blei ist schwerer als **Papier** (T) Eis ist **heiß** (F) Häuser haben eine **Tür** (T) Blumen brauchen **Licht** (T)
Set 2 Menschen putzen ihre Zähne mit einem **Löffel** (F)	**Set 2** Kinder mögen **Eis** (T) Schweine können **fliegen** (F)	**Set 5** Menschen schlafen in einem **Bett** (T) Möhren/Karotten können **tanzen** (F) Äpfel wachsen am **Baum** (T)	**Set 8** Häuser sind aus **Käse** (F) Katzen mögen **schlafen** (T) Fleischer machen **Brot** (F) Worte bilden einen **Satz** (T)	**Set 11** Menschen haben eine **Nase** (T) Saft enthält viel **Fett** (F) Eine Rose ist ein **Tier** (F) Eine Maus ist kleiner als ein **Hund** (T) Ein Auto kann **fahren** (T)
Set 3 Es gibt Gras im **Park** (T)	**Set 3** Die Erde hat einen **Mond** (T) Papier kann **kochen** (F)	**Set 6** Deutschland hat einen **König** (F) Kühe essen gerne **Gras** (T) Ein Kapitän steuert ein **Schiff** (T)	**Set 9** Hasen können **lesen** (F) Hühner essen **Holz** (F) Kinder gehen in die **Schule** (T) Ein Zug fährt auf einem **See** (F)	**Set 12** Ein Schuh hat einen **Kopf** (F) Pferde rennen im **Himmel** (F) Eine Uhr zeigt die **Zeit** (T) Ein Buch kann **laufen** (F) Ein Lachs ist ein **Fisch** (T)

Note. Target items are highlighted with bold; T = True; F = False.
Target items were controlled for frequency and length (high frequency, one- or two-syllable long words). The task included only simple sentences and it was modified to avoid semantic and phonological interference across items to ensure the highest recall rate possible. Each stimulus set of the task was checked by a native German speaker for phonological and semantic similarity to avoid interference. Immediately after hearing each sentence, participants were asked to judge it as true or false by pointing a check mark or cross on a sheet of paper. Concurrently, they were asked to retain the final word of each sentence in each set for spoken recall, immediately after the entire set was presented. Probe sets were included at the level of 1 and 2, having performed before presenting the experimental trials.

Level 1 Probe set 1: Menschen lesen Bücher im **Ofen** (F)
Probe set 2: Hunde haben einen **Schwanz** (T)
Level 2 Probe set 3: Die Zwiebel ist ein **Obst** (F); Ein Elefant hat einen **Rüssel** (T)
Probe set 4: Menschen sehen Löwen im **Zoo** (W); Zwölf ist gleich **Duzend** (W)

Table S4. Raw data obtained in the outcome measures.

	Case					
	E.Q.		I.B.		M.N.	
Outcome measure	Pre-post	FU	Pre-post	FU	Pre-post	FU
N-back with letters						
1-back (max. 30)	26-24		29-30		22-26	
2-back (max. 30)	15-23		17-25		12-7	
3-back (max. 30)	5-22		13-21		8-5	
Running span (max. 30)	16-20		5-8		6-8	
TROG-D (max. 84)	65-62	68	64-68	70	45-59	59
Token test (max. 50)	25-29		30-30		12-16	
Sätze verstehen						
Short irreversible (max. 22)	22-21		21-22		17-18	
Long irreversible (max. 22)	22-22		19-20		19-16	
Case-marked SVO (max. 20)	20-16		19-20		13-13	
Case-marked OVS (max. 20)	3-6		3-0		10-9	
Number-marked SVO (max. 20)	17-14		17-20		15-13	
Number-marked OVS (max. 20)	2-13		3-2		8-12	
Right-branching SRC (max. 20)	13-13		16-19		9-11	
Right-branching ORC (max. 20)	5-5		3-2		5-6	
Centre-embedded SRC (max. 20)	9-8		10-13		7-8	
Centre-embedded ORC (max. 20)	2-4		1-2		2-5	
Total (max. 204)	115-122		112-120		105-111	
Canonical (max. 80)	59-51		62-72		44-45	
Non-canonical (max. 80)	12-28		10-6		25-32	
ANELT						
Understandability	2.6-3		2.4-3.3		1.4-1.5	
Number of words	20-25.6		32.7-30.3		11.5-22.8	
CIUs	15.44-20.33		18.66-20.7		4.9-9.2	
%CIUs	79.73-84.56		62.03-76.43		44.76-46.73	
CIUs/min	65.03-70.2		58.2-76.13		25.91-26.32	
Words/min	78.57-79.24		98.18-100.24		54.11-60.27	
EMQ						
Speech (max. 4)	1.62-0.77		0.54-0.23		1.38-1.46	
Learning new things (max. 4)	-		-		1.5-1	